January/February 2004 · Vol. 6 · No. 1

An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions

Muin J. Khoury, MD, PhD[1], Quanhe Yang, PhD[2], Marta Gwinn, MD, MS[1], Julian Little, PhD[1,3], W. Dana Flanders, MD, DSc[4]

Purpose: The current clinical value of genomic profiling (testing for genotypes at multiple loci) for assessing susceptibility to common diseases and targeting behavioral and medical interventions is questionable. As common diseases result from many gene-environment interactions, epidemiologic studies should be used to examine the value of genomic profiling in terms of clinical validity (future disease positive and negative predictive value stratified by exposure), clinical utility (targeted interventions to reduce disease risk among persons with the profile) and public health utility (comparing reduction of disease burden in the population based on genomic profiling to population-wide interventions). **Methods:** We investigate these parameters for a hypothetical common disease (5% lifetime risk), for which 3 genetic variants at different loci and one environmental exposure are risk factors. **Results:** We show that even for modest effects of each variant alone (risk ratios from 1.5–3.0) and modest interactions between the exposure and the genes, the disease predictive value for people with 2 or more variants (especially 3) can be quite high (50–100%) in the presence of a modifiable exposure. Individual risks can then be reduced by targeted exposure intervention among persons with the genotype. However, the predictive value for multiple genotypes is much lower for rarer diseases ($<$ 1 per 1000). Also, with increasing number of genes in a profile, the population impact of disease reduction for targeted intervention based on genotype will be smaller, especially for rare genotypes, weak associations, and weak interactions. **Conclusion:** To assess the value of genomic profiling, well-designed epidemiologic studies are needed to quantify disease risks, in addition to costs, benefits, and risks for testing and interventions. ***Genet Med* 2004:6(1):38–47.**

Key Words: *epidemiology, genetic testing, gene-environment interaction, genomic profiling*

The completion of the human genome sequence is an important milestone for biology, health, and society.[1–3] Advances in genomics could play a central role in practice by providing genetic information for disease prediction and prevention. In 2001, Collins and McKusick[3] predicted: By the year 2010, it is expected that predictive genetic tests will be available for as many as a dozen common conditions, allowing individuals who wish to know this information to learn their individual susceptibilities and to take steps to reduce those risks for which interventions are or will be available. Such interventions could take the form of medical surveillance, lifestyle modifications, diet, or drug therapy. Identification of persons at highest risk for colon cancer, for example, could lead to targeted efforts to provide colonoscopic screening to those individuals, with the likelihood of preventing many premature deaths.

Nevertheless, several authors have expressed skepticism regarding the value of susceptibility genetic testing for disease prevention.[4–9] Concerns cited include the absence of interventions that are specific to different genotypes, the potential detraction from the proven benefits of population-wide prevention based on known risk factors (smoking, diet, and physical activity), the ethical, legal, and social ramifications of genetic information, and the low magnitude of risk for common diseases associated with most genetic variants discovered thus far. In contrast to genotypes for single gene disorders such as Huntington disease, most genotypes for common complex diseases are incompletely penetrant because of the interaction with other genotypes and environmental factors at large (diet, drugs, and infectious agents), and genotype-phenotype correlations will tend to be weak, leading to uncertainties about the meaning of positive and negative genetic tests.[4,8] For example, Holtzman and Marteau[4] showed that the positive predictive value for common disease genotypes (i.e., probability that a person with a specific genotype will develop disease) tends to

From the [1]Office of Genomics and Disease Prevention, [2]Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia; [3]Department of Epidemiology, University of Aberdeen, Aberdeen, Scotland; [4]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia.

Dr. Muin J. Khoury, Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road, MS E82, Atlanta, GA 303033.

Received: August 18, 2003.

Accepted: October 18, 2003.

DOI: 10.1097/01.GIM.0000105751.71430.79

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

be low unless the genotype is rare (<1%) or the relative risks are high (20 or more) even for relatively common diseases with lifetime risks of 5%. When those lifetime risks in a given population are lower, the positive predictive value will also be much lower. Because the disease predictive ability of tests for common variants at single loci is low, such variants seem unsuitable for use in clinical practice. In addition, associations between a disease and a genotype can be observed as a result of bias, chance, or publication bias.[10]

Despite the cautionary notes about the value of common genetic polymorphisms as a basis for predicting disease in the future and targeting interventions, currently a number of companies in the United States and the United Kingdom are offering testing for multiple genetic polymorphisms as part of genomic profiling for susceptibility to various conditions including obesity, cardiovascular disease, and susceptibility to infectious diseases and autoimmunity.[11,12] Although testing for common genetic polymorphisms is currently not ready for clinical practice,[13–14] we need to anticipate that any future use of such testing should be based on several types of objective data to validate its clinical validity and utility.[15,16]

An important consideration for genetic susceptibility testing is the concept of using genetic variants at multiple loci (referred to as genomic profiling throughout this article) that individually are weak risk factors for a complex disease but collectively may better predict future disease. Yang et al.[17] showed that bundling several variants from multiple loci that interact in one or more biological pathway (e.g., folate metabolic pathways) could increase the predictive value of genetic testing for susceptibility to common disease, especially in the presence of pertinent environmental exposures (e.g., dietary and supplemental folic acid intake). Combining several genetic variants from different loci that fall in one or more biological pathway for specific exposures makes biological and clinical sense. Examples include the many variants in the cytochrome P450 genes interacting with drugs and environmental toxicants,[18] genetic variation in thrombosis cascade genes interacting with hormonal therapy,[19] and variants in folate metabolism genes interacting with dietary or supplemental folate intake.[20]

In this article, we explore the epidemiologic building blocks for assessing the potential use of genomic profiling to predict common diseases. We show how the concept of gene-environment interaction can be used to define the clinical relevance of genetic testing at multiple loci for common disease susceptibility. Although our illustrations are limited, they can be generalized to more genes and more modifiable risk factors. We show how the measurement in epidemiologic studies of disease associations with multiple genotypes and environmental factors has an effect on the clinical validity and utility of genetic susceptibility tests. For common chronic diseases, under certain conditions, genomic profiling may lead to high disease predictive values, in the presence of common interacting factors such as drugs and diet. The value of genomic profiling for achieving prevention goals should be based on epidemiologic parameters of risk as well as on laboratory, economic, ethical, legal, and social considerations of testing and targeting interventions. In this genomic era, we will increasingly compare the relative merits of population-based and high-risk approaches (based on genotypes) in preventing disease and improving health in the population.

METHODS

We explore what happens in a hypothetical population with a known lifetime risk of a common disease (see Appendix). We use 5% to correspond to a common condition such as colorectal cancer[21] (we also consider rarer disease conditions of 5 per 10,000, corresponding to relatively common birth defects such as cleft palate[22]). We assume that underlying this lifetime risk are the joint effects of measured genetic variants at three unlinked loci, and one modifiable risk factor or exposure, along with other unmeasured factors. We use the term "exposure" broadly to include modifiable factors that increase disease risk (such as cigarette smoking) as well as factors or procedures that decrease risk (such as medical procedures or chemo preventive agents). For simplicity, we assume a dichotomous susceptibility genotype at each locus and also a dichotomous exposure. In reality, many more loci and exposures are involved. We assume that the effects of additional genes and exposures are not directly measured here as part of the risk characterization equations. Also, we do not deal in this study with the uncertainties in measuring exposures or with the effects of dose, timing, and duration of exposures.

The combination of genotypes and exposure, their corresponding frequencies in the population and conditional lifetime risks for disease can be arranged in a multilevel table (see Table 5 in the appendix). Because epidemiologic studies typically quantify associations between diseases and genotypes or exposures in terms of relative risks (or odds ratios), Table 5 in the appendix lays out risk estimates in terms of risk ratios for each genotype at one locus, the exposure, and joint effects of exposures and one or more genotypes. The appendix also displays all the assumptions behind the analyses presented. This basic model can be extended to reflect multiple genes and exposures.

Epidemiologic assessment of clinical validity and utility for testing for multiple genotypes

The clinical validity and utility of a genetic test have been defined by the task force on genetic testing[23] and by the Secretary's Advisory Committee on Genetic Testing.[15] In the context of testing for multiple genetic variants for susceptibility to future disease occurrence, the crucial parameters are the positive and negative predictive values of the test. The predictive value is the probability of future disease given a combination of genotypes and an interacting modifiable exposure (see Table 6 in the appendix). This table also provides a framework for assessing the clinical utility of the test. Clinical utility reflects the ability to lower disease risks for people with a "positive" genetic test. In this case scenario, because of targeted interventions on the modifiable risk factors (which could be a chemp-

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

roventive drug, lifestyle modification, diet change, early detection of disease), we can assess the reduction in risk for individuals with 0, 1+, 2+, and 3 variants, if we intervene on the environmental side by reducing the risk from E+ to E−.

In addition to clinical validity and utility, SACGT recommended that the analytic validity of the test (sensitivity and specificity to measure correct genotypes) be high, and also for the ethical, legal, and social implications (ELSI) of the test be appropriately assessed.[15] In this study, we assume an analytically valid test although this cannot be taken for granted in practice as test developers start bundling up multiple gene variants in the same assay. These issues are briefly alluded to in the discussion section.

Epidemiologic assessment of public health utility for testing for multiple genotypes

Another way to assess the value of genetic testing for multiple genes is the public health utility, or population impact of this approach. We compare two forms of interventions with respect to their effect on disease risk reduction in the population. The first is an intervention directed at the whole population irrespective of genotype to remove the exposure. The second is an intervention targeted to the high-risk group on the basis of risk stratification by genotype followed. We compare the relative impact of each approach by quantifying the values of population attributable fraction of disease (PAF) to assess how much disease burden in the population is associated with specific exposure and therefore what would be the reduction in disease burden if the exposure were removed.[24] Feigelson et al. (H. Feigelson, American Cancer Society, personal communication, 2003) derived values of exposure population attributable fractions for targeted interventions (PAFt) based on genotype. In this study, we extend their analysis of one gene and one exposure to 3 genotypes and one exposure. Using the formulas in the appendix, we can derive the ratio of PAFt to PAF. The closer the ratio is to unity, the closer the impact of targeted intervention based on genetic testing will be to a population approach for exposure reduction.

RESULTS

Because many combinations of relevant parameters are possible, we limit our displays to a few examples (Tables 1–4). We also discuss below the impact of changing some of these variables. In Tables 1 and 2, we illustrate the situation for a common disease with lifetime risk of 5% in the population, three relatively uncommon genotypes at three loci (each with 5% population frequency) with modest association with disease risk (we vary risk ratio from 1.5 to 3, which is typical in many epidemiologic studies). We assume the exposure risk ratio also varies from 1.5 to 3. We also assume the joint effect between one exposure and one genotype has a modest increase over the product of their individual relative risks (by a factor of 1.5, often called synergy index; see appendix for further detail). As shown in Table 1, 14% of the population will have one or more susceptibility genotypes whereas a very small fraction (0.7%) has 2 or more. Also, the disease predictive values stratified by exposure and genotypes are highest for 3 genotypes (+exposures), increase with increasing risk ratios, and are generally higher for less common exposures (5% vs. 50%). For some combinations, predictive values in the presence of an exposure are quite high in the range of 50% to 100%, similar to many

Table 1

Disease predictive values (%) stratified by genotype and environmental exposure, by risk ratios of individual genes and exposures, and exposure frequency (for a disease with population lifetime risk of 5%, S = 1.5) and rare genotypes (5%)[a]

	Genomic profile							
	0		1+		2+		3	
	85.7		14.3		0.7		0.01	
Population % exposure	No	Yes	No	Yes	No	Yes	No	Yes
Exp Frequency 5%								
Risk ratio								
1.5	3.3	5.0	6.9	16.2	10.2	34.8	15.2	76.8
2.0	2.0	4.0	8.5	26.9	16.5	75.5	32.5	100
3.0	1.0	3.0	11.0	53.4	30.9	100	89.5	100
Exp Frequency 50%								
Risk ratio								
1.5	2.0	3.0	5.4	12.5	7.9	27.0	11.7	59.4
2.0	1.5	3.0	5.5	17.4	10.7	48.8	20.9	100
3.0	0.7	2.0	5.4	26.0	15.0	100	43.6	100

[a]Genomic profile refers to the number of "susceptibility" genotypes at the 3 independent loci that a person has. The value ranges from 0 to 3.
S = 1.5 refers to the multiplicative synergy index for joint effects of an exposure and a genotype (see appendix for details).
Predictive values of more than 100% represent combination of parameters not mathematically plausible.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

Table 2
Population impact (measured in terms of population attributable fraction) of targeted environmental interventions based on genotype compared to a general reduction in exposure in the population, by risk ratios of individual genes and exposures and exposure frequency (for a disease with population lifetime risk of 5%, S = 1.5) and rare genotypes (5%)[a]

		Targeted to genomic profile[b]		
		1+	2+	3
Intervention Population %	Population-wide	14.3	0.7	0.01
Exp frequency 5%				
Risk Ratio				
1.5	3.2	1.3 (40.7%)	0.2 (5.6%)	0.01 (0.3%)
2.0	6.1	2.6 (42.6%)	0.4 (6.6%)	0.02 (0.3%)
3.0	11.7	6.0 (51.3%)	1.3 (11.1%)	0.1 (0.9%)
Exp frequency 50%				
Risk Ratio				
1.5	25.1	10.2 (40.7%)	1.4 (5.5%)	0.06 (0.2%)
2.0	39.4	16.9 (42.6%)	2.8 (7.1%)	0.2 (0.5%)
3.0	57.1	29.4 (51.5%)	6.5 (11.4%)	0.5 (0.9%)

[a]Genomic profile refers to the number of "susceptibility" genotypes at the 3 independent loci that a person has. The value ranges from 0 to 3.
[b]Attributable Fractions % (AFt/AFpop %).
S = 1.5 refers to the multiplicative synergy index for joint effects of an exposure and a genotype (see appendix for details).

Table 3
Disease predictive values (%) stratified by genotype and environmental exposure, by risk ratios of individual genes and exposures and exposure frequency (for a disease with population lifetime risk of 5%, S = 1.5) and common genotypes (50%)[a]

	Genomic profile							
	0		1+		2+		3	
	12.5		87.5		50.0		12.5	
Population % exposure	No	Yes	No	Yes	No	Yes	No	Yes
Exp Frequency 5%								
Risk ratio								
1.5	1.3	2.0	4.8	16.4	5.8	22.9	7.8	39.2
2.0	0.5	1.0	4.6	22.5	6.2	33.7	10.0	67.4
3.0	0.001	0.005	4.2	33.1	6.3	53.1	12.6	100
Exp Frequency 50%								
Risk ratio								
1.5	0.7	1.0	2.5	8.5	3.0	11.9	4.0	20.4
2.0	0.5	1.0	1.9	9.3	2.6	13.9	4.1	27.8
3.0	0.0003	0.001	1.3	10.1	1.9	16.2	3.8	38.8

[a]Genomic profile refers to the number of "susceptibility" genotypes at the 3 independent loci that a person has. The value ranges from 0 to 3.
S = 1.5 refers to the multiplicative synergy index for joint effects of an exposure and a genotype (see appendix for details).
Predictive values of more than 100% represent combination of parameters not mathematically plausible.

single-gene disorders with incomplete penetrance (e.g., *BRCA1* in breast/ovarian cancer[25]). For example, with risk ratios of 2, the lifetime predictive value for people with 2+ genotypes is 75.5% for exposed people and 16.5% for unexposed people. These hypothetical data not only provide an estimate of the potential clinical validity of such testing but also the potential amount of risk reduction to the individual with susceptibility genotypes on the basis of exposure reduction. (i.e., clinical utility).

What about the population impact of such profiling? Table 2 shows the effects on overall disease risk in the population based on the scenario presented in Table 1. For rare exposures (5%), a population-wide intervention regardless of genotype will reduce the disease burden by only 3% to 11.7% (depending on

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

Table 4

Population impact (measured in terms of population attributable fraction) of targeted environmental interventions based on genotype compared to a general reduction in exposure in the population, by risk ratios of individual genes and exposures and exposure frequency (for a disease with population lifetime risk of 5%, S = 1.5) and common genotypes (50%)[a]

		Targeted to genomic profile[b]		
		1+	2+	3
Intervention population %	Population-wide	87.5	50.0	12.5
Exp frequency 5%				
Risk ratio				
1.5	10.3	10.2 (98.5%)	8.5 (82.9%)	3.9 (38.3%)
2.0	15.8	15.6 (98.7%)	13.7 (86.7%)	7.2 (45.6%)
3.0	25.4	25.3 (99.6%)	23.4 (92.1%)	14.4 (56.7%)
Exp frequency 50%				
Risk ratio				
1.5	53.4	52.7 (98.6%)	44.3 (82.9%)	20.4 (38.2%)
2.0	65.2	64.5 (98.9%)	56.8 (87.1%)	29.7 (45.6%)
3.0	77.3	76.9 (99.5%)	71.3 (92.2%)	43.8 (56.7%)

[a]Genomic profile refers to the number of "susceptibility" genotypes at the 3 independent loci that a person has. The value ranges from 0 to 3.
[b]Attributable Fractions % (AFt/AFpop %).
S = 1.5 refers to the multiplicative synergy index for joint effects of an exposure and a genotype (see appendix for details).

the risk ratios). Such a scenario would make the exposure an unlikely candidate for a public health intervention. Moreover, targeting interventions by genotype will lead to a smaller reduction in the disease burden in the population primarily because the prevalence of multiple genotypes is low. When the exposure is more common (50% prevalence), the population-wide intervention will lead to a more substantial decline in the disease burden (by 25% to 57% depending on risk ratios). We can also see that by targeting 14% of the population with one or more genotypes, we can achieve a 40% to 51% reduction in the overall disease burden related to exposure in the population (depending on risk ratios). To summarize, in the presence of relatively uncommon genotypes, although individual predictive values can become quite high among persons with susceptibility genotypes and exposures, the overall population impact of risk reduction is generally low if environmental risk reduction is based on genotype.

Tables 3 and 4 show analyses similar to Tables 1 and 2 with one exception. The genotype frequencies are now much higher (50%). As shown, 87.5% of the population has one or more genotypes, 50% have two or more, and 12.5% have three susceptibility genotypes. In this case, the predictive values shown in Table 3 are generally lower than those presented in Table 1. However, for some combinations of parameters, they can be in the high range of 50% to 100%. For example, among the 12.5% of the population with three genotypes, for risk ratios of 2, the disease predictive value is 67.4% in the presence of exposure and 10% in the absence of exposure. Substantial reduction in disease risk can be achieved in this group by removal of the interacting environmental factor. How will this translate into population impact or public health utility? Table 4 shows the same analysis presented in Table 2 but for common genotypes. Here, we can see that the overall reduction in disease burden based on population intervention regardless of genotype will be more substantial (varies from 10% to 77%). However, we can also see that targeting interventions to people with genotypes can achieve most of this overall reduction. For example, with risk ratios of 2, the ratio of targeted attributable fraction over total attributable fraction is about 98% for one or more genotype, 86% for two or more variants, and 45% for three variants. In other words, by targeting environmental risk reduction to 12.5% of the population with 3 susceptibility genotypes, we can prevent almost half of the burden of disease in the population due to the exposure.

What happens when some of the parameters in these tables are varied? The most important parameter for disease predictive value is overall disease risk in the population. Let us look at the numbers in Tables 1 and 3 and apply them to a disease with an overall population risk of 5 per 10,000 instead of 5 per 100. This scenario corresponds to many relatively common birth defects for example (such as cleft palate). All the predictive values in the tables need to be divided by 100, which makes the clinical validity and utility of such testing essentially not suitable in clinical practice. This is because the risk ratios reflect weak to moderate effects (1.5 to 3). Although not shown, the predictive will increase substantially if individual risk ratios are in the range of 10 to 30. Therefore, the clinical usefulness for genomic profiling for common genetic risk factors with modest effects is only relevant to common diseases. On the other hand, if the overall population risk is more than 5% (say 10% or 25%), which may be relevant to common conditions like hypertension and coronary heart disease, the predictive values presented in Tables 1 and 3 become higher with the same risk ratios (1.5 to 3).

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

Another parameter is the impact of joint effects of genotypes and exposures. The true biological forms for joint effects for most genes and exposures are not well known but could vary from additive, multiplicative, or supramultiplicative joint effects. Although not shown, in general, the higher the synergy index, the higher the predictive values will be as well as the population impact of targeted interventions. These values will be lower for pure additive effects between genotypes and exposures. Although not shown here, the existence of epistasis across multiple gene loci (i.e., gene-gene interactions) will lead to higher predictive values of genomic profiling compared to those measured without such interaction.

Finally, let us examine the situation when more than three genes are included in a profile. This last scenario will be more likely to occur in practice in the next decade or two. Although no analyses are shown in this study, if other parameters are kept constant, the predictive values will increase with increasing the numbers of genotypes (regardless of the underlying model of interaction) but the population impact of such intervention based on genotypes will diminish rapidly because the intervention will apply to a smaller and smaller fraction of the population.

DISCUSSION

With the completion of the Human Genome Project, there is an increasing expectation that finding genetic variation associated with susceptibility to common diseases (e.g., cancer and coronary heart disease) will lead to the development of susceptibility genetic tests that predict the risk of future disease and lead to targeted interventions.[1,2] These interventions could include primary prevention (such as dietary changes and physical activity), secondary prevention (such as early detection of cancer through biochemical and radiological tests), and tertiary prevention and therapeutics (such as targeted pharmacological agents). The premise is that by using information from genetic variants and their products at multiple loci, we will be able to construct "genomic profiles" of risks for various diseases. Such a scenario is reflected in some of the futuristic predictions for the practice of medicine.[2] In the future, genomic profiles may contain dozens or hundreds of genetic variants. They could also be based on gene-expression profiles,[26] or protein expression variation.[27]

Currently, there are no obvious applications of genomic profiling that we can use in disease prevention despite the increasing offering of such genomic profiles by commercial entities.[28] It is becoming clear from the emerging published literature on gene-disease associations and gene-environment interaction that many such associations are not replicated because they could be due to chance or limitations in study design.[29] Moreover, as indicated above, even if replicated across studies or in metaanalyses (e.g., the relationship between *MTHFR* and the risk of coronary heart disease[30]) such associations may not have immediate clinical applications for targeting interventions. One important reason is that for complex common diseases with multiple risk factors assessing genetic risk factors one at a time or the joint effects of one genetic factor and one environmental factor at a time has so far led us to only weak or modest associations in terms of risk ratios and population attributable fractions.

In this article, we apply a simple epidemiologic framework to the assessment of the validity and utility of testing for multiple genetic variants for susceptibility for future occurrence of common diseases. Using the epidemiologic concepts of risk ratios and population attributable fractions, even with the limited examples shown, it is clear that the predictive value for testing one gene at a time, even in the presence of an interacting exposure, will lead to relatively low predictive values even for a common disease with lifetime risk of 5%. However, we also show that by increasing the numbers of variants in a "genetic test" the picture will change for common diseases. Even for relatively weak associations between individual genotypes and a disease (risk ratios of 1.5 to 3), which is typical of many modern association studies, and also with modest synergistic joint effects between exposures and genotypes, the magnitudes of predictive values for exposed individuals can be quite high under some circumstances. With only three genes, we could approach predictive values in the range of 50% to 100%, which is equivalent to a single gene disorder with incomplete penetrance (e.g., *BRCA1* mutations and lifetime risk of breast or ovarian cancer). Predictive values could be made much higher by increasing the number of gene variants in a genomic profile or risk. Future work needs to explore the joint effects of variation in numerous genes (hundreds and may be thousands) on the predictive values for different disease outcomes. In this regard, methodological and statistical work is still emerging (e.g., recent combinatorial partitioning methods to assess multiple variable loci for quantitative traits[31])

Interestingly, for a common disease, the overall effect on disease predictive values and risk reduction of combining several common genetic variants at different loci is similar to the effect of combining several environmental risk factors (or protective factors). The recent analysis of the "polypill" concept illustrates this point. Wald and Law[32] proposed that a formulation that contains low-dose aspirin, folic acid, low-dose statin, and blood pressure-lowering drugs can reduce the incidence of ischemic heart disease by 88% and stroke by 80%. Similar to our analysis, combining several factors that individually have modest effects on disease risk and prevention can have a profound overall effect on disease risk (or risk reduction) in the face of a common disease such as heart disease and stroke.

It is noteworthy that for most epidemiologic studies, relative risks measured in various studies usually refer to ratios in incidence rates (person-years analysis) or cumulative disease rates over short periods of time (e.g., 5–10 years).[24] Therefore, even for common diseases with lifetime risks of 5% to 10%, absolute risks will still be relatively low when measured using incidence rates or short-term cumulative risks. On the other hand, geneticists normally think in terms of penetrance (lifetime risks of disease) in relation to genotypes. Therefore, for common diseases, it is not difficult to see how relatively weak

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

or moderate associations of risk ratios in the range of 2 when examined in relation to multiple genes can lead to high lifetime risks (i.e., penetrance). It is also important to recognize the underlying assumptions for using risk ratios to derive lifetime risks (i.e., no competing risks and stable incidence rates over time and age cohorts[24]), although it is possible to account for competing risks, e.g., using life table analyses, but the calculations become more complex (illustrated by breast cancer[33]).

This article illustrates the well-known tension between "high-risk" and "population" approaches to prevention of common diseases.[34] The population approach to reducing risk factor prevalence in the whole population, if successful, will lead to maximum benefits of prevention of diseases associated with these risk factors, whereas targeting high-risk individuals on the basis of a genomic profile for risk reduction could miss a substantial fraction of disease in the population. Nevertheless, despite our knowledge about primary risk factors for many chronic diseases, we continue to face tremendous challenges in implementing population-wide approaches to risk reduction of common diseases, e.g., through messages about smoking cessation, diet, exercise, and adherence to recommended medical interventions. For example, more than 60% of people do not get enough physical activity,[35] 23% of the United States population still smokes cigarettes,[36] 21% of people are obese,[37] and only 44% adhere to recommendations related to colorectal cancer screening.[38] With the advent of genomics, the ability to increasingly target such intervention to people who need them the most will force us to assess carefully the "value-added" of such an approach.

Without detracting from population messages on prevention, some segments of the population could benefit from a more intensive approach to achieve prevention goals. A balancing act between high-risk and population intervention should be based on risk characterization. To the individual, knowledge of lifetime risks (predictive values) with and without an intervention, could lead to substantial risk reduction under some circumstances discussed above. These are the concepts of clinical validity and utility promoted by recent advisory groups.[15,23] We show in this study how these concepts are directly related to epidemiologic measures of risk and risk reduction based on exposure modification by genotype. In addition, we extend the concept of clinical utility to public health utility based on the work of Feigelson et al. (H. Feigelson, personal communication, 2003), which directly compares the benefits of disease reduction in the population by a general exposure reduction independent of genotype versus disease reduction based on targeted intervention to high-risk genotypes. By using the ratio of two attributable fractions, we can estimate how much of the overall disease reduction can be achieved by initial targeting before interventions. As expected, with more genes added to a genomic profile, the individual predictive values for disease risk will be higher but the population impact of targeted intervention will be lower, because fewer people will have the combination of the "susceptible" genetic variants. Therefore, whereas such interventions may make sense at an individual level, they may not make sense on a population level, if the objective is to achieve maximum reduction in the burden of disease in the population.

Although a population approach may be superior when dealing with exposures or practices that are known to account for a large attributable fraction of the disease, such as cigarette smoking and lung cancer (population attributable fraction of $> 90\%$), it may not be the case when the overall contribution of the exposure to disease occurrence is much less. The illustrations in Tables 2 and 4 show this point. For example, in Table 2, we see that for an exposure frequency of 5% and risk ratios of 2, the population attributable fraction of the exposure is only 6%. By targeting interventions to the 14% of the population that have one or more genes, we can achieve a 43% reduction in the proportion of disease associated with this exposure. In Table 4, we show that for an exposure frequency of 5% and risk ratios of only 1.5, the overall exposure attributable fraction is 10% but we will be able to achieve an 83% reduction in that risk by targeting the 50% of the population with two or more genotypes. In other words, there could be instances that a population-wide approach may not be done because the risk factor is not an important overall cause of disease. The "high risk" approach (targeted by genotype) may uncover subsets of the population with markedly increased risk, thereby meriting more aggressive individualized interventions.[39]

In the final analysis, both a high-risk approach (based on genotype) and a population approach could be needed to achieve prevention goals for individuals and populations. In his classic discussion of this subject, Rose concluded: "If causes can be removed, susceptibility ceases to matter. Realistically, many diseases will long continue to call for both (population and high-risk) approaches, and fortunately competition between them is usually unnecessary."[34] The combination of population and high-risk approaches for chronic disease prevention has been recently highlighted by Hunt et al.[40] in the context of coronary heart disease prevention. They show that screening the general population for family history of heart disease can combine the benefits of population-wide education with more intensive assessment directed only to a high-risk subset (e.g., to diagnose familial hypercholesterolemia), because most heart disease events, especially those that occur at an early age, are concentrated in a small fraction of families.[40]

One example of a successful population approach that targets interventions to only a small fraction of persons at risk is newborn screening.[41] Newborn screening programs have prevented unnecessary mental retardation due to Phenylketonuria and congenital hypothyroidism and deaths from conditions such as sickle cell disease, despite the relative rarity of these conditions. In such a scenario, the magnitude of gene-environment is so extreme (the effect of the exposure is only limited to people with the genotype[42]) that only targeted intervention that starts with population-wide search makes sense. However, in the majority of complex diseases with numerous gene-environment interactions, it is highly unlikely that one exposure or one gene could provide the magic bullets for interventions.

It is important to consider some of the potential limitations of the analyses presented here. First, risk characterization

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

should be based on properly conducted epidemiologic studies of genotype-disease associations and gene-environment interaction. Because of the observational nature of these studies, issues around unbiased selection of subjects, sample size, biologic plausibility, adjustment for potential confounders (including population stratification), and replication of findings across populations all need to be adequately considered.[10] Second, there are emerging challenges for how to consider simultaneously the impact of several, potentially hundreds if not thousands of genetic variants to arrive at valid estimates of disease risk .[10] New methods will have to be explored to assess multiple comparisons, using analytic methods based on a prior biological models and analysis of joint effects of genes and exposures. Third, the illustrations presented in this study are limited in scope and have multiple assumptions as described in the appendix. Empiric data derived from epidemiologic studies of multiples genes and exposures, are needed to quantify biological interactions. For example, a recent epidemiologic study investigated the interaction between polymorphisms in the *CHEK2* gene and *BRCA1/2* mutations in relation to the risk of breast cancer.[43] The study showed that a *CHEK2* truncating variant is associated with a relative risk of 2 for breast cancer only among women without *BRCA1/2* mutations, but no increase in risk for *BRCA1/2* mutations carriers. Although the results need further confirmation, they suggest that joint effects of gene products on the same biological pathway may not always be synergistic in the way shown in this article but that the effects of one gene product may be subsumed under another major gene on the same pathway.[43] In any case, epidemiologic data are needed for real populations.

Another limitation is that we discuss only one special situation in which individual genotypes have equal population frequencies and independent effects on disease risks. In reality, for many diseases, some genes have major influences on risks, whereas others may have a lesser role in disease occurrence.

Finally, epidemiologic estimates of risks that are derived from empirical data are only a first step to arrive at appropriate clinical or population guidelines. Issues not considered here include the costs of testing, the analytic performance of tests, the types and costs of interventions, the timing of testing in relation to the natural history of disease, the psychosocial impact of testing and interventions, and the potential for stigmatization from labeling people as susceptible. These issues have usually been included in principles of population screening, which may have to be reevaluated in this genomics era.[41] An additional important consideration for clinical utility for the individual is whether or not knowledge of increased risk on the basis of a genomic profile will enhance adoption of medical and behavioral interventions that reduce disease risk. Conversely, individuals without a "positive" genomic profile may still be at increased disease risk from the exposure, albeit at lower levels, and thus may become more complacent about not engaging in healthy behaviors or seeking appropriate preventive interventions.

In conclusion, genomic profiling for measuring susceptibility to common diseases and targeting medical and behavioral interventions can be assessed using epidemiologic studies that estimate the magnitudes of relative, absolute, and attributable risks. Although genomic profiling is not likely to be ready for clinical use for some time, it is important to consider that genetic variants with weak-to-modest associations with common diseases (i.e., relative risks of 1.5–3) will have limited clinical value for predicting disease susceptibility if used alone. However, for common genetic variants, especially for those that may interact in one or more defined biological pathway(s), measuring their combined effects on disease risk (along with exposures) may hold promise in increasing the value of such variants as part of an overall risk profile. In order to fulfill the promise of the Human Genome Project, well-conducted epidemiologic studies are now urgently needed to assess the added value of genomic profiling for preventing disease and improving health in the 21st century.

References

1. Collins FS, Green ED, Guttmacher AE et al. A vision for the future of genomics research. *Nature* 2003;422:835–847.
2. Collins FS. Shattuck lecture: Medical and societal consequences of the Human Genome Project. *N Engl J Med* 1999;341:28–37.
3. Collins FS, McKusick VA. Implications of the Human Genome Project for medical science. *JAMA* 2001;285:540–544.
4. Holtzman NA, Marteau TM. Will genetics revolutionize medicine? *N Engl J Med* 2000;343:141–144.
5. Zimmern RL. The human genome project: a false dawn? *BMJ* 1999;319:1282.
6. Cooper RS, Psaty BM. Genomics and medicine: distraction, incremental progress, or the dawn of a new age? *Ann Intern Med* 2003;138:576–580.
7. Willett WC. Balancing life-style and genomics research for disease prevention. *Science* 2002;296:695–698.
8. Evans JP, Skrzynia C, Burke W. The complexities of predictive genetic testing. *BMJ* 2001;322:1052–1056.
9. Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. *Lancet* 2001;357:709–712.
10. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol* 2002;156:300–310.
11. Sciona. Discover the relationship between you and your genes. Available at: http://www.sciona.com. Accessed June 1, 2003.
12. Genovations: the advent of truly personalized healthcare. Available at: http://www.genovations.com. Accessed June 1, 2003.
13. Haga S, Khoury MJ, Burke W. Genomic profiling for lifestyle modification: not ready for prime time. *Nat Genet* 2003;34:347–350.
14. Barrett S, Hall H. Dubious genetic testing: Quackwatch. Available at: http://www.quackwatch.org/01QuackeryRelatedTopics/Tests/genomics.html. Accessed August 5, 2003.
15. Secretary's Advisory Committee on Genetic Testing (SACGT). Enhancing the Oversight of Genetic Tests: Recommendations of the SACGT (2000).Available at: http://www4.od.nih.gov/oba/sacgt/reports/oversight_report.htm. Accessed online, June 2, 2003.
16. Burke W, Atkins A, Gwinn M et al. Genetic test evaluations: information needs for clinicians, policy makers and the public. *Am J Epidemiol* 2002;156:311–318.
17. Yang Q, Khoury MJ, Botto L et al. Improving the prediction of complex diseases by testing for multiple disease susceptibility genes. *Am J Hum Genet* 2003;72:636–649.
18. Nebert DW, Russell DW. Clinical importance of the cytochromes P450. *Lancet* 2002;9340:1155–1162.
19. Rosendaal FR, Helmerhorst FM, Vandenbroucke JP. Female hormones and thrombosis. *Arterioscler Thromb Vasc Biol* 2002;22:201–210.
20. Stover PJ, Garza C. Bringing individuality to public health recommendations. *J Nutr* 2002;132(suppl):2476S–2480S.
21. National Cancer Institute. Probability of Developing or Dying of Cancer. Database available at: http://srab.cancer.gov/devcan/canques.html. Accessed June 2, 2003.
22. Mossey PA, Little J. Epidemiology of oral clefts: an international perspective. In: Wyszynski DF, editor. Cleft lip and palate: From origins to treatment. Oxford University Press; 2002:127–158.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

23. Holtzman NA, Waston MS, editors. Promoting safe and effective use of genetic testing: Final Report of the Task Force on Genetic Testing, 1997. Available at: http://www.genome.gov/page.cfm?pageID=10001733. Accessed June 2, 2003.
24. Khoury MJ, Beaty TH, Cohen BH. Fundamentals of genetic epidemiology. New York: Oxford University Press; 1993:77–79.
25. Burke W, Austin M. Genetic risk in context: calculating the penetrance of *BRCA1* and *BRCA2* mutations. *J Natl Cancer Inst* 2002;94:1185–1187.
26. van de Vijver MJ, He YD, van 't Veer LJ et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med* 2002;347:1999–2009.
27. Petricoin EF, Ardekani AM, Hitt BA et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;359:572–579
28. Khoury MJ. Genetics and genomics in practice: the continuum from genetic disease to genetic information in health and disease. *Genet Med* 2003;5:261–268.
29. Little J, Khoury MJ, Bradley L et al. The Human genome is complete. How do we develop a handle for the pump? *Am J Epidemiol* 2003;157:667–673.
30. Klerk M, Verhoef V, Clarke R et al. *MTHFR* 677C→T Polymorphism and risk of coronary heart disease. A meta analysis. *JAMA* 2002;288:2023–2031.
31. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;3:458–470.
32. Wald NJ, Law MR. A strategy to reduce cardiovascular risk by more than 80%. *BMJ* 2003;326:1419–1424.
33. Feuer EJ, Wun LM, Boring CC et al. The lifetime risk of developing breast cancer. *JNCI* 1993;85:892–897.
34. Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985;14:32–38.
35. Centers for Disease Control, and Prevention. Physical activity and good nutrition: essential elements to prevent chronic diseases and obesity. At-a-glance 2002. website accessed January 1, 2003 Available at: http://www.cdc.gov/nccdphp/aag/aag_dnpa.htm
36. Centers for Disease Control, and Prevention. Current cigarette smoking among adults aged 18 and older, 2000. Available at: http://www.cdc.gov/tobacco/statehi/html_2002/current_2000.htm. Website January 1, 2003.
37. Mokdad AH, Ford ES, Bowman BA et al. Prevalence of obesity, diabetes and obesity-related health risk factors, 2001. *JAMA* 2003;289:76–79.
38. Centers for Disease Control, and Prevention. Trends in Screening for Colorectal Cancer: United States,. and 1999. MMWR 2001;50:162–166. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5009a2.htm, 1997. Accessed January 1, 2003.
39. Rockhill B. The privatization of risk. *Am J Public Health* 2001;91:365–368.
40. Hunt S, Gwinn M, Adams TD. Family history assessment: strategies for prevention of cardiovascular disease. *Am J Prev Med* 2003;24:136–142.
41. Khoury MJ, McCabe L, McCabe ERB. Population screening in the age of genomic medicine. *N Engl J Med* 2003;348:50–58
42. Khoury MJ, Adams MJ, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet* 1988;42:89–95.
43. Meijers-Heijboer H, van den Ouweland A, Wasielewski M, et al. Low penetrance susceptibility to breast cancer due to *CHECK2* (*)1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat Genet* 2002;31:55–59.

Appendix

Epidemiologic assessment of lifetime risks for a disease based on three genes and one exposure, by exposure and genotype frequencies, risk ratios, and joint effects

For a disease with a lifetime risk D in the population, we consider three independent dichotomous disease susceptibility genotypes with one dichotomous environmental exposure risk factor. The population can be partitioned into 16 strata depending on the combination of disease susceptibility genes and environmental exposure (see Table 5). We define P_{ijkl} to be the proportion of the population with gene 1 = *i*, gene 2 = *j*, gene 3 = *k*, exposure = *l* and *i*,*j*,*k*,*l*, = 0,1. RR_{ijkl} is the risk at each individual combination of genotype and exposure relative to risk at $i = j = k = l = 0$. For example, P_{1001} is the proportion of the population with disease susceptibility gene 1 = 1, gene 2 and gene 3 = 0 and exposed ($l = 1$), and RR_{1011} is the risk ratio for the disease among those who carry disease susceptibility genes 1 and 3 and are exposed compared with those who are carrying no disease susceptibility genes and are not exposed.

To simplify the situation for illustration, we assume the following: (1) independence of the distribution of the three genes and the exposure in the population; (2) each gene variant has the same prevalence and risk ratios; (3) Synergy for joint effects occurs only between exposures and genes, but not among genes.

We use the following nomenclature: G is the population prevalence of the susceptibility genotype at each locus (0, vari-

Table 5
Population frequencies and disease risks stratified by genotypes and exposure status[a]

Gene 1 (i)	Gene 2 (j)	Gene 3 (k)	Exposure (l)	Pop freq P_{ijkl}	Risk
0	0	0	0	$(1-G)^3 \cdot (1-E)$	I
0	0	0	1	$(1-G)^3 \cdot E$	IR_e
0	0	1	0	$G(1-G)^2(1-E)$	IR_g
0	0	1	1	$G(1-G)^2E$	IR_gR_eS
0	1	0	0	$G(1-G)^2(1-E)$	IR_g
0	1	0	1	$G(1-G)^2E$	IR_gR_eS
1	0	0	0	$G(1-G)^2(1-E)$	IR_g
1	0	0	1	$G(1-G)^2E$	IR_gR_eS
0	1	1	0	$G^2(1-G)(1-E)$	IR_gR_g
0	1	1	1	$G^2(1-G)E$	$IR_gR_gR_eS^2$
1	0	1	0	$G^2(1-G)(1-E)$	IR_gR_g
1	0	1	1	$G^2(1-G)E$	$IR_gR_gR_eS^2$
1	1	0	0	$G^2(1-G)(1-E)$	IR_gR_g
1	1	0	1	$G^2(1-G)E$	$IR_gR_gR_eS^2$
1	1	1	0	$G^3(1-E)$	$IR_gR_gR_g$
1	1	1	1	G^3E	$IR_gR_gR_gR_eS^3$

[a]Where I is background lifetime risk of disease in the absence of the three genotypes and the exposure. We choose values for R_g, R_eS, and the P_{ijkl}, then set $\Sigma\Sigma\Sigma\Sigma\ P_{ijkl}(RR_{ijkl})I = 0.05$ and solve for *I* to satisfy this assumption. We restrict the estimations to the range $0 < R_{ijkl}{}^*I < 1$.

Table 6
Notations for disease predictive values based on a combination of genotypes and an interacting modifiable exposure[a]

Genotype	Exposure −	Exposure +
NPV 0	P(D/0, E−)	P(D/0, E+)
PPV 1 (or more)	P(D/1+, E−)	P(D/1+, E+)
PPV 2 (or more)	P(D/2+.E−)	P(D/2+, E+)
PPV 3	P(D/3, E−)	P(D/3, E+)

[a]NPV is negative predictive value and PPV is positive predictive value): Only the first row (genotype = 0) can be viewed as the test negative predictive value. All the others reflect different levels of positive predictive values (depending on whether the positive genetic test is defined as carrying, 1 or more, 2 or more, or 3 variants). This can be extended to more genes. Because Table 5 has all the elements of disease risks by strata, values of predictive values in the table above can be easily computed based on the quantities in Table 5.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.

ant absent and 1, variant present); E = population prevalence of exposure (0, absent; 1, present); R_g = lifetime risk ratio for disease for genotype 1 compared to 0 (at one locus); and R_e = the lifetime risk ratio for disease for exposure 1 compared to exposure 0.

For both R_g and R_e, we assume no confounding and competing risks. S is synergy index for combined effects of genotype and exposure (if S = 1, it implies multiplicative effects of risk ratio). We can then fill the multiway contingency table (Table 5).

We also assume the following model for the risk,

$$R_{(i,j,k,l)} = I \cdot R_{g1}^{i} R_{g2}^{j} R_{g3}^{k} R_{e}^{l} S^{[(g1+g2+g3)e]} \tag{1}$$

Disease predictive value among exposed and unexposed

Disease predictive value among exposed is defined as the life time probability of developing the disease in people with the variant genotypes and exposed.

$$P(D/\vec{G}, E = 1) = \frac{I \sum_i \sum_j \sum_k P_{ijk1} R_{ijk1}}{\sum_i \sum_j \sum_k P_{ijk1}} \tag{2}$$

where I is the background life time risk of disease in the absence of the genotypes and the exposure. The disease predictive value among the unexposed is defined as the life time probability of developing the disease in people with the variant genotypes but absence of the exposure.

$$P(D/\vec{G}, E = 0) = \frac{I \sum_i \sum_j \sum_k P_{ijk0} R_{ijk0}}{\sum_i \sum_j \sum_k P_{ijk0}} \tag{3}$$

Population attributable fractions with three genes and one environmental exposure

The following expression is for the population attributable fraction for exposure, assuming no-confounding[24]:

$$AF_{pop} = \frac{P_e(RR - 1)}{1 + P_e(RR - 1)} \tag{4}$$

where P_e is the proportion of the population that is exposed to environmental risk factor (E in Table 5). *RR* is the risk ratio of exposure for the disease. The above formula assumes that everyone in the population is equally susceptible for the disease. When considering three independent dichotomous disease susceptibility genotypes and a dichotomous environmental exposure risk factor, the attributable fraction becomes the following:

$$AF_{pop} = \frac{\begin{array}{c} P_{0001}(RR_{0001} - 1) + P_{1001}(RR_{1001} - RR_{1000}) \\ + P_{0101}(RR_{0101} - RR_{0100}) + P_{0011}(RR_{0011} - RR_{0010}) \\ + P_{1101}(RR_{1101} - RR_{1100}) + P_{1011}(RR_{1011} - RR_{1010}) \\ + P_{0111}(RR_{0111} - RR_{0110}) + P_{1111}(RR_{1111} - RR_{1110} \end{array}}{\sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} \sum_{l=0}^{1} P_{ijkl}(RR_{ijkl})} \tag{5}$$

under the given assumptions, the values of these parameters can be obtained from Table 5. The numerator of this equation includes a contribution from each component of the population that is exposed across all genotypes. The denominator is the overall population risk for the disease [P(D), which we have fixed at 5% in our illustration]. This equation measures the overall reduction of disease risk in the population if the exposure is removed in the population.

The targeted population attributable fraction (H. Feigelson, American Cancer Society, personal communication, 2003), AF_T, estimates the maximum potential impact on the population of a targeted intervention designed to eliminate the exposure among those who carry a susceptibility genotype or combination of multiple genotypes and exposure. For example, if we are to target interventions to persons with one or more genotypes,AF_T/AF_{POP} represents the ratio of the two attributable fractions. The closer this is to unity, the closer the overall impact of targeted intervention will be to the overall reduction of disease if exposure is removed from the population.

Copyright © American College of Medical Genetics. Unauthorized reproduction of this article is prohibited.